WHAT MAKES A CUSTOMER BUY A COSMETIC PRODUCT?

Amanda Sanelisiwe Nkomo
Mario Alcaraz

BDA 620: DATA MINING

Table of Contents

INTRODUCTION

EXECUTIVE SUMMARY

Customer Behavior has a significant impact on the success of e-commerce, and we need to understand customer behavior to optimize online transactions and increase the number of overall purchases. “As in the case of traditional marketing in the past, most of the recent research and debate is focused on the identification and analysis of factors that one way or another can influence or even shape the online consumer’s behaviour “(Miles et al., 2000, p131). Customer Behaviour can be assessed by looking at the online buying process and the events that take place in the buying decision processes. Looking into how online users’ follow through with transactions and the factors that influence the users to make the purchase of the product. Common factors could be the price of the product, the brand name, product reviews, complexities of the online portals and product awareness.

Electronic Commerce, also known as E-commerce, is the buying and selling of goods and services on online platforms with the use of electronic transactions. This is usually done on online platforms, websites, and mobile applications. An Event type refers to a specific action or behavior that a user may take during online transactions to gain insights about user behavior and preferences. Events are used to gather data about user behaviour, and this data is then used to improve the user experience, optimize marketing campaigns, and increase conversions (Efthymios Constantinides,2004).

“Customer Lifetime Value (CLV) is a marketing metric that projects the value of a customer over the entire history of that customer's relationship with a company” (Ramachandran,2006). There are several factors that can impact CLV in e-commerce, including the types of products or services being offered, the frequency of purchases, the average order value, and customer retention rates. By understanding these factors and tracking CLV over time, e-commerce businesses can identify areas for improvement and take proactive steps to increase customer value and drive growth.

Customer journey is the various touchpoints and interactions a customer has with a brand or business across different stages of the buying process, from initial awareness to post-purchase activities. A positive customer journey and experience can lead to customer loyalty and repeat business, while a negative experience can result in customer dissatisfaction, attrition, and negative feedback.

BUSINESS PROBLEM

The objective of the study is to analyze purchase prediction and factors that may influence customer behavior, using the case of an E-commerce Cosmetic Store (October 2019). By analyzing the event-type data, businesses can gain valuable insights into customer behaviour and preferences, which can be used to optimize marketing strategies and increase sales.

BUSINESS QUESTIONS

1. What factors influence customers' decision to make a purchase?
2. How can the business reduce the occurrence of customers removing products from their carts?
3. What is the customer lifetime value (CLV) of different customer segments, and how can marketing efforts be optimized to increase CLV for E commerce business?

DATA EXPLORATION AND PROCESSING

DATA DESCRIPTION AND DICTIONARY

E-commerce Cosmetic Store Oct 2019, Source: REES46 Marketing Platform ,4,102,283 Observations and 9 variables .The dataset contains behavioural data for a medium cosmetics online store.

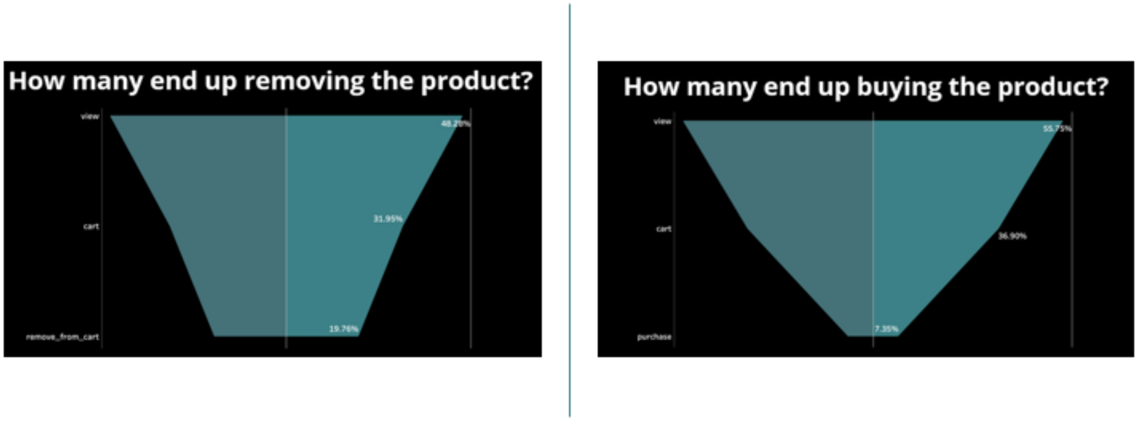
event time **(String - Coordinated Universal Time)**	Time when event happened at (in UTC).
event type ***(Categorical)***	View - a user viewed a product Cart - a user added a product to shopping cart Remove_from_cart - a user removed a product from shopping cart Purchase - a user purchased a product
product_id ***(Integer)***	ID of a product
category_id ***(Numerical)***	Product's category ID
category code ***(categorical)***	Product's category code
Brand ***(Categorical)***	String of brand names.
Price ***(Numerical)***	Float price of a product.
user_id ***(Integer)***	Permanent user ID.
user session ***(integer)***	Temporary user's session ID. Changes every time user come back to online store from a long pause.

Table.1: E-Commerce Cosmetic Store October 2019 Data Dictionary

Missing Values

- category code 4,034,806 missing observations
- brand 1,659,261 missing observations
- user session 637 missing observations

DATA EXPLORATION



(Fig. 1) Event Type Analysis

From the chart above the event types, we note that view, cart and remove from cart are all bigger proportions of the dataset as compared to purchases which have a minority/smaller number of instances. This shows that our dataset is imbalanced and what we want to predict(purchases) is minority class.

Price Distribution

The price distribution is left side/ negatively skewed, which could indicate that many prices are low. Also, by looking into the summary statistics the price range seems to be lower than $10 for the products with the most frequent price being crucial information. The price distribution and box plot have been attached to Appendix B.

Products

After running an analysis of product frequency, we were able to identify that some products are more common than others, indicating popularity.

Brands

Found 240 different brands among the dataset. However, it was clear that not all of them have the same number of products. This led to the question, Are some brands bigger than others? Does this matter for a customer?

DATA CLEANING

Dimension Reduction

The first step taken for dimension reduction, reducing the number of observations. The reason behind doing this is computing power. While having more data helps to train a better model, it comes with computing power being required and in the case of a business more money spent. Therefore, a sample of 1.2M observations was created.

(Note: the sample size was constantly adjusted to meet the desired minimum number of records after balancing the dataset)

Due to the nature of the variables holding an 'ID' identifier, they were dropped and not considered for model building. The reason behind this decision is that an id containing a numerical value higher than another does not mean it has more importance or more significant. Hence, the coefficient would be irrelevant. Additionally, the variable 'category code' was also removed due to the high number of missing values and insignificant amount of information provided. The following variables were dropped from the dataset:

- Cartegory_ID
- User_ID
- User_Session
- Category_Code
- Product_ID (Held for the creation of new variable before dropping it)

Data Type Transformation

- **Event Time**: The variable event time had a datatype of 'object'. Therefore, a transformation to datatype date type was performed.

New Variables Created

From our exploratory data analysis, we identified some product characteristics that would make it more susceptible to being viewed, added to the cart a purchased. Some products were more frequent than others. Also, some brands had a bigger amount of distinct product_id, indicating more variety or a brand of larger scale. However, we could not directly use id or brand due to the large number of variations.

The solution to this:

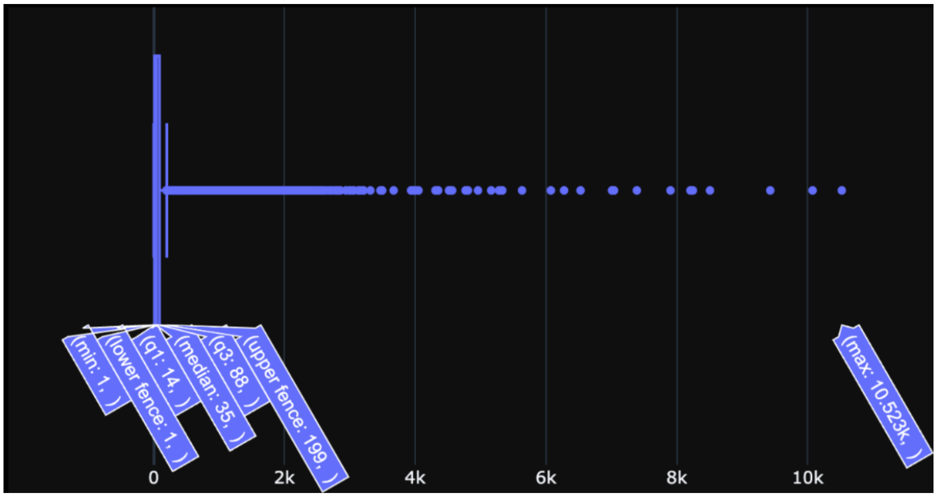
Big Brand Variable

- Count unique products per brand.
- Determine what makes it a big brand.
- Apply conditional function to choose 1 or 0 (1 = Big Brand, 0 = Not Big Brand).

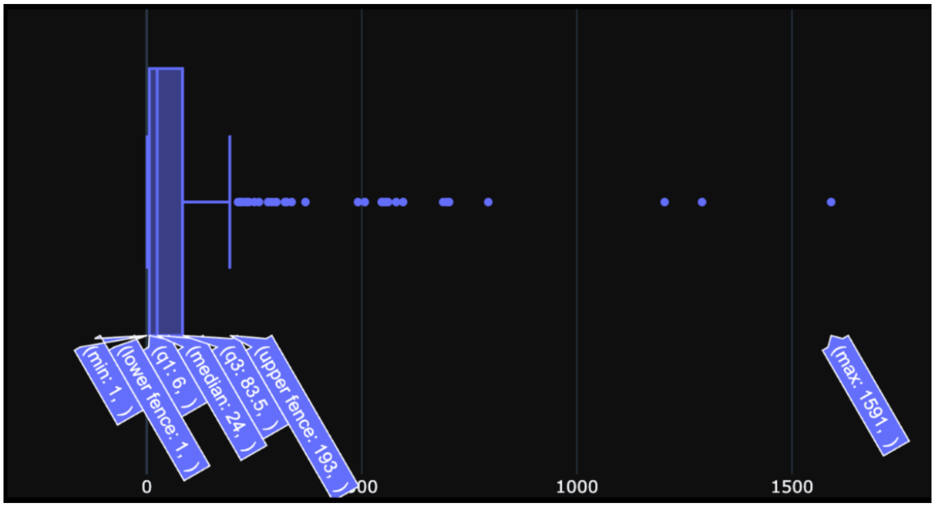
Popular product Variable.

- Count the number of appearances for each product.
- Determine what makes it a popular product
- Apply conditional function to choose 1 or 0 (1 = Popular Product, 0 = Not Popular).

To determine the cutoff of what makes it a popular product or a big brand, we decided to take a statistical approach. A box plot of the distribution of both (Fig. 1, Fig. 2) shows the upper fence of the distribution. Any number from a brand or product above the respective upper fence would be considered as popular or big.



(Fig. 2) Product Popularity Distribution



(Fig. 3) Brand Size Distribution

Data Partitioning

Training dataset: This data set is constituted by 60% of the data sample

Validation dataset: This dataset is constituted by the remaining 40% of the data sample. Used to validate model results.

Data Balancing

A supervised model learns based on what you feed it. Having a dataset with clean variables was not sufficient to run a model. The dataset still had a disproportionate number of observations belonging to different event types. This would inevitably cause a model that predicts better a specific event type.

Random Over-Sampling Examples (ROSE) is a technique used to balance minority classes by randomly selecting observations to oversample and balance the data to an even or close to even number for the classes. This package was used to balance our training dataset.

MODEL SIMULATION AND RESULTS

LOGISTIC REGRESSION

Logistic regression models were used to analyze the relationship between binary dependent variables (event type) and the predictor variables (Big brand, popular brand, price, day, hour). The model coefficients and odds ratios quantify the strength of the relationship between the independent variables and the dependent variable. A confusion matrix can be used to evaluate the performance of the logistic regression model. Overall, logistic regression is a powerful tool for analyzing customer behavior and predicting the likelihood of a customer making a purchase.

Model 1: View to Cart

Model coefficients are statistically significant and negatively associated with increasing the likelihood that a product would be added to the cart.

Confusion Matrix: Cart

		PREDICTED CLASS	
		Cart “1”	No Cart “0”
ACTUAL CLASS	Cart “1”	14512	25953
	No Cart “0”	15376	44159

Accuracy: 58%

Sensitivity: 62%

Specificity: 48%

Model 2: Cart to Purchase

Model coefficients are statistically significant and big brand, price, and hour are negatively associated with increasing the likelihood that a product would be added to the cart, while popular product and day are positively associated with increasing the likelihood that a product would be purchased.

Confusion Matrix: Purchase

		PREDICTED CLASS	
		Purchase “1”	No Purchase “0”
ACTUAL CLASS	Purchase “1”	928	9930
	No Purchase “0”	4917	84225

Accuracy: 85%

Sensitivity: 89%

Specificity: 15%

Model 3: Cart to Remove from Cart

Model coefficients are statistically significant and popular product, price, and hour are negatively associated with increasing the likelihood that a product would be removed from the cart, while big brand and day are positively associated with increasing the likelihood that a product would be purchased.

Confusion Matrix: Remove from Cart

		PREDICTED CLASS	
		Removed “1”	Not Removed “0”
ACTUAL CLASS	Removed “1”	5774	18785
	Not Removed “0”	12520	62921

Accuracy: 68%

Sensitivity:77%

Specificity: 31%

NAIVE BAYES

Naïve bayes was an additional technique used. Different than other models, this technique is more of a statistical model based on conditional probabilities. The model gets prior probabilities of the class and the features. Then, it takes the probabilities from an observation and calculates P(A/B). Given this, what is the probability of that? Data balancing and proper sampling was particularly important for this model. Having more data would lead to more observations of a characteristic, increasing its probability and affecting results.

Model 1: View to Cart

		PREDICTED CLASS	
		Cart “1”	No Cart “0”
ACTUAL CLASS	Cart “1”	114703	242536
	No Cart “0”	28961	93800

Accuracy: 43%

Sensitivity: 27%

Specificity: 79%

Model 2: Cart to Purchase

		PREDICTED CLASS	
		Purchase “1”	No Purchase “0”
ACTUAL CLASS	Purchase “1”	27,285	270,704
	No Purchase “0”	984	181,027

Accuracy: 433%

Sensitivity: 40%

Specificity: 96%

Model 3: Cart to Remove from Cart

		PREDICTED CLASS	
		Removed “1”	Not Removed “0”
ACTUAL CLASS	Removed “1”	78,006	186,977
	Not Removed “0”	10,065	204,952

Accuracy: 58%

Sensitivity:52%

Specificity: 88%

The results obtained from this model had a wide variation. Adjusting to obtain more specificity had a big negative impact on accuracy. The mode did a fair job on prior iterations for sensitivity to the cost of incredibly low specificity. For this reason, we opted to focus on improving the logistic models.

CONCLUSION

In comparison to the Naïve Bayes, the Logistic Regression models seem to be able to make predictions with reasonable accuracy, but there is room for improvement, especially in terms of specificity. We have attached the lift and decile charts in Appendix B. We on adopting the Logistic regression Model, modeling the prediction of making a purchase of a cosmetic product for further analysis.

Limitations we faced were dealing with the observations and having to sample the dataset, information about the user id and the user sessions were lost. This would be relevant for future analysis, particularly relating the customer lifetime value and purchase predictions.

RECOMMENDATIONS

Analysing the data on the products that customers are adding to their cart or purchasing can help businesses make personalized recommendations to customers. This can increase the likelihood of a repeat purchase and enhance customer experience.

By promoting popular products, the business can Identify the most popular products, and prioritize them in marketing campaigns. Using social media, email marketing, and other channels to promote these products and highlight their unique features and benefits. The business can run promotions on popular days of the week or times of the year when the customers are most likely to make a purchase and run promotions or special offers during these times. This can help to increase sales and drive customer loyalty.

In the context of e-commerce and with the use of the logistic regression models to predict purchasing outcomes, CLV (Customer Lifetime Value) can measure the potential long-term revenue that a customer can generate for an online business. Considering various factors such as the number of purchases made by the customer, the average order value, and the length of time the customer has been making purchases from the business. The model can help to identify those customers who make a purchase, and

CLV can be used to identify the most valuable customers and focus marketing efforts on retaining them. It can also help to identify opportunities to increase revenue from existing customers by offering them personalized promotions, product recommendations, or loyalty rewards.

To curb the loss of sales from abandoned carts and products removed from the cart the business can, send cart abandonment notifications when a customer removes items from their cart and does not complete their purchase. The business can also simplify the checkout process and make sure the checkout process is simple and user-friendly. Customers are more likely to complete a purchase if the checkout process is streamlined and easy to navigate. The business may also implement Retargeting advertisements and displaying them to users on other websites or social media platforms, the business can encourage them to return and complete their purchase.

APPENDIX

APPENDIX A:

REFERENCES:

I. Miles, G.E., Howes, A. and Davis, A. (2000), “A framework for understanding human factors in Web-based electronic commerce”, International Journal of Human-Computer Studies, Vol. 52 No. 1, pp. 131-63.

II. Efthymios Constantinides, Influencing the online consumer’s behaviour: the Web experience, Volume 14 · Number 2 · 2004 · 111-126.

III. Ming Zeng1 & Hancheng Cao1 & Min Chen2 & Yong Li1, July 2017, User behaviour modelling, recommendations, and purchase prediction during shopping festivals

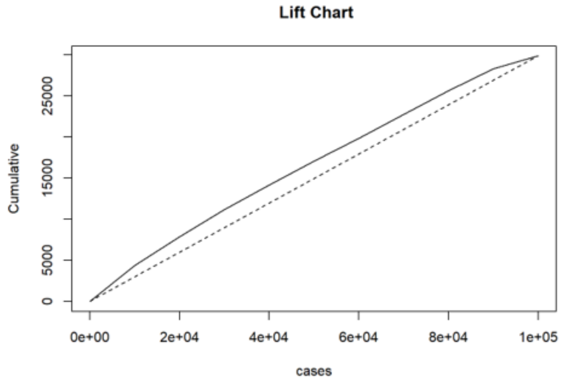
IV. https://www.turing.com/kb/an-introduction-to-naive-bayes-algorithm-for-beginners

V. Customer Lifetime Value,Ramakrishnan Ramachandran,February 2006, DOI:10.13140/2.1.1787.1049, Conference: National Seminar On "Changing Scenario of Consumerism" organized by Department of Commerce, Bharathidasan University Tiruchirapalli, India

VI. https://rees46.com/en/datasets

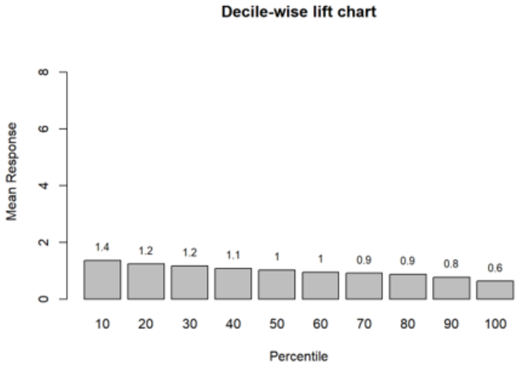
APPENDIX B:

Purchase Model

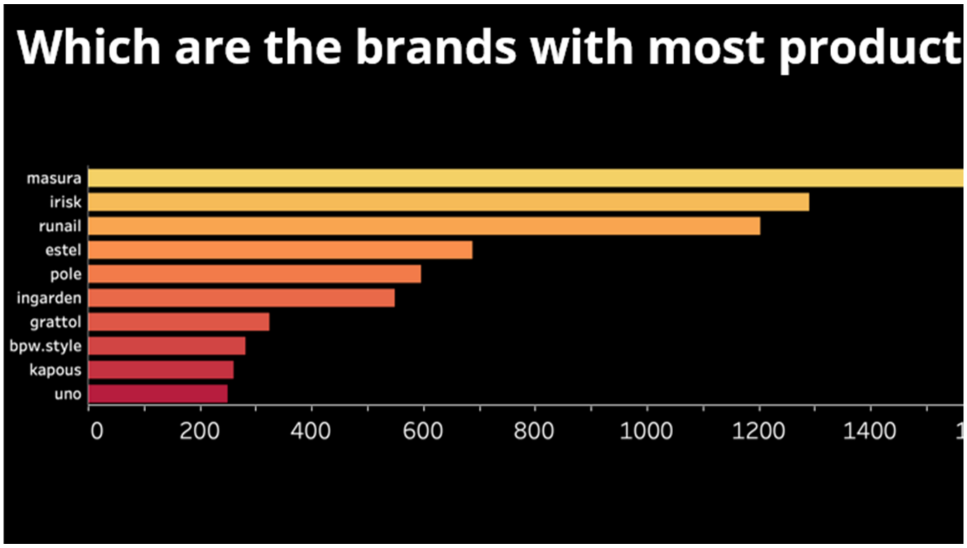


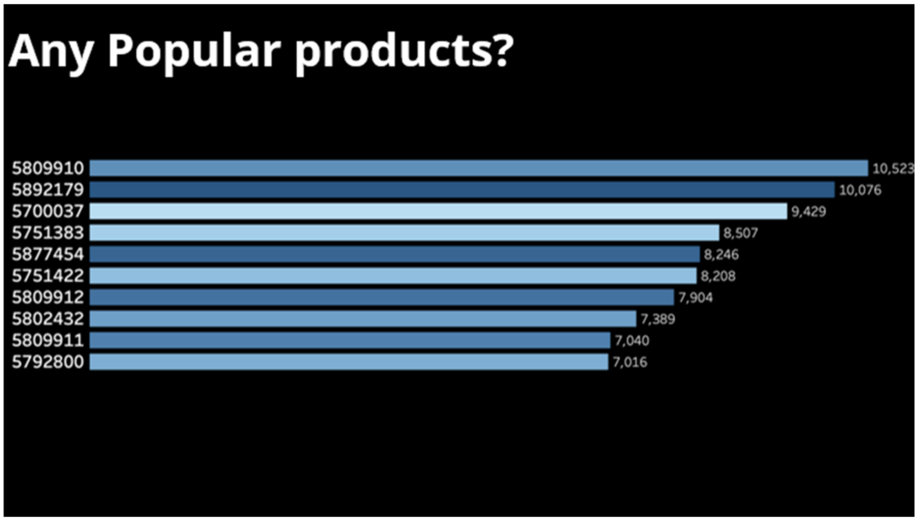
Lift Chart

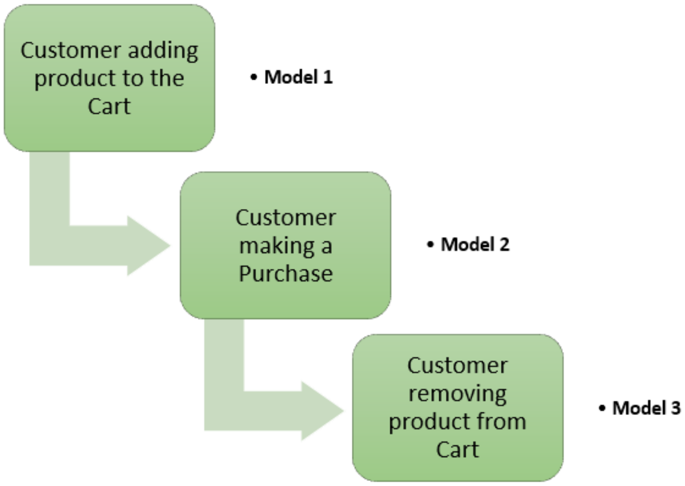
Purchase Model



Decile-wise lift chart







Customer adding product to the Cart

- **Model 1**

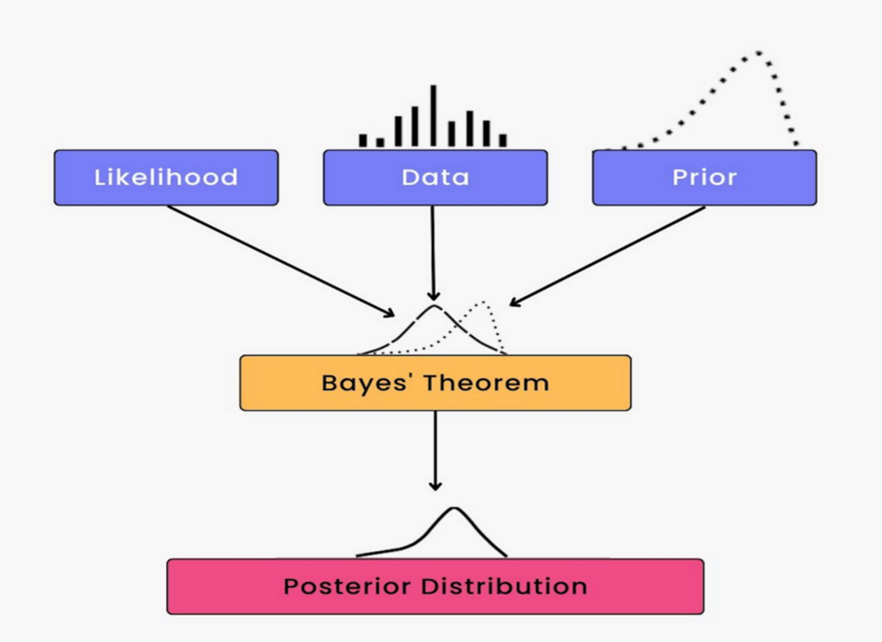
Customer making a Purchase

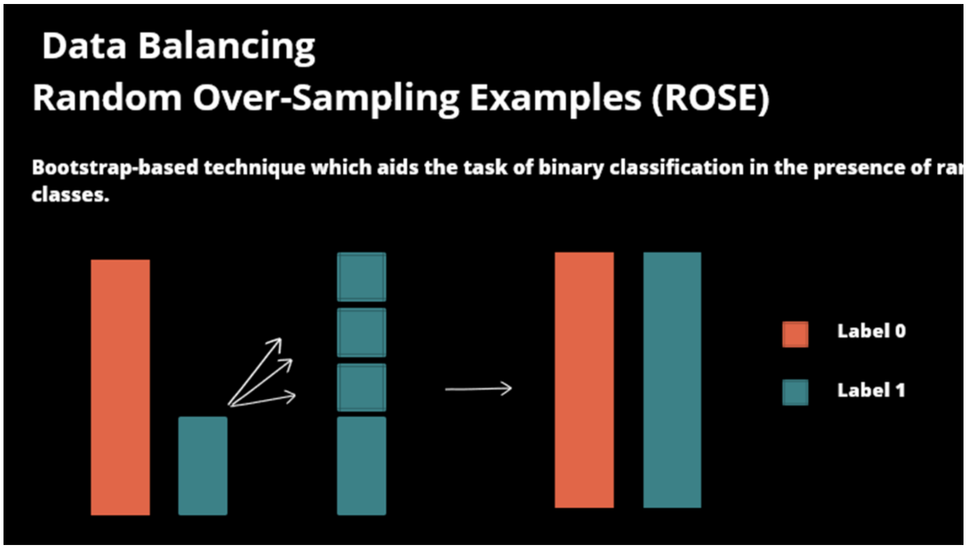
- **Model 2**

Customer removing product from Cart

- **Model 3**

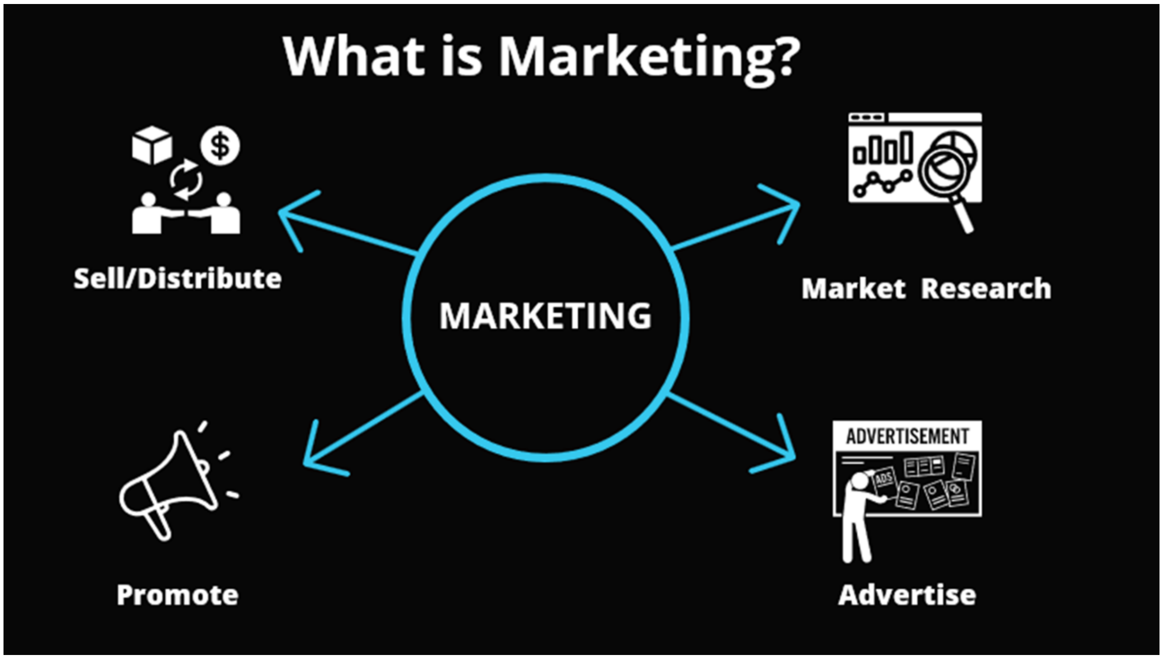
\





Data Balancing
Random Over-Sampling Examples (ROSE)

Bootstrap-based technique which aids the task of binary classification in the presence of ra classes.



What is Marketing?

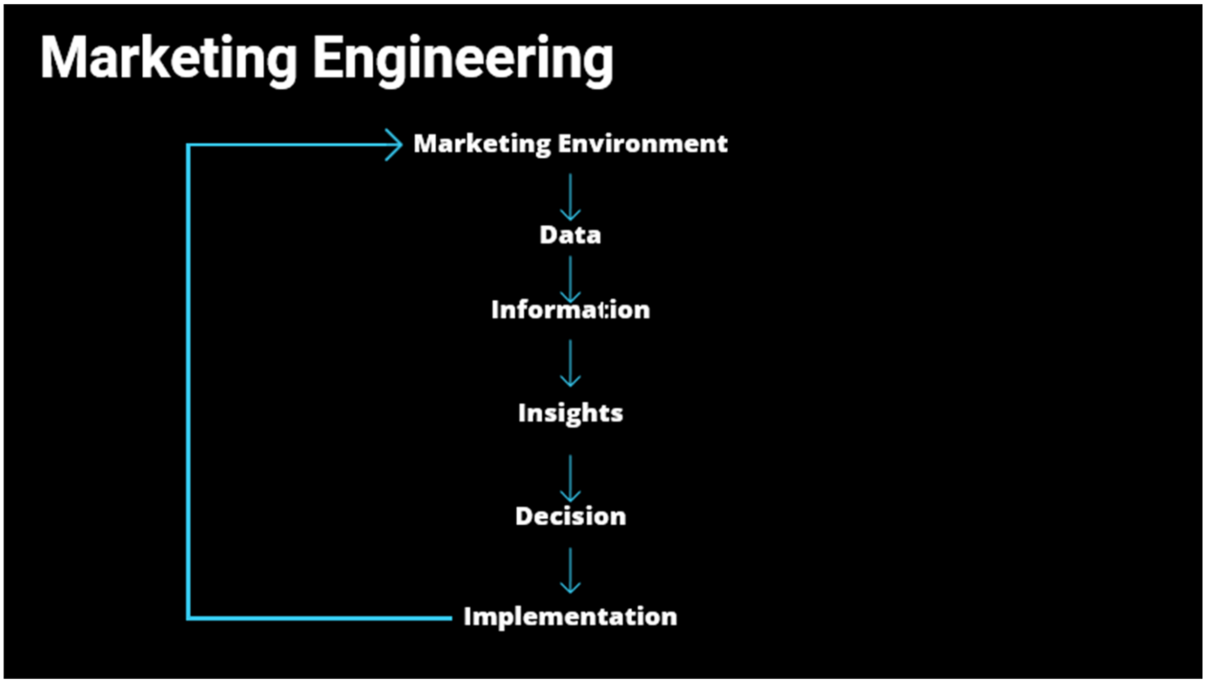
MARKETING

Sell/Distribute

Market Research

Promote

Advertise



Marketing Engineering

Marketing Environment

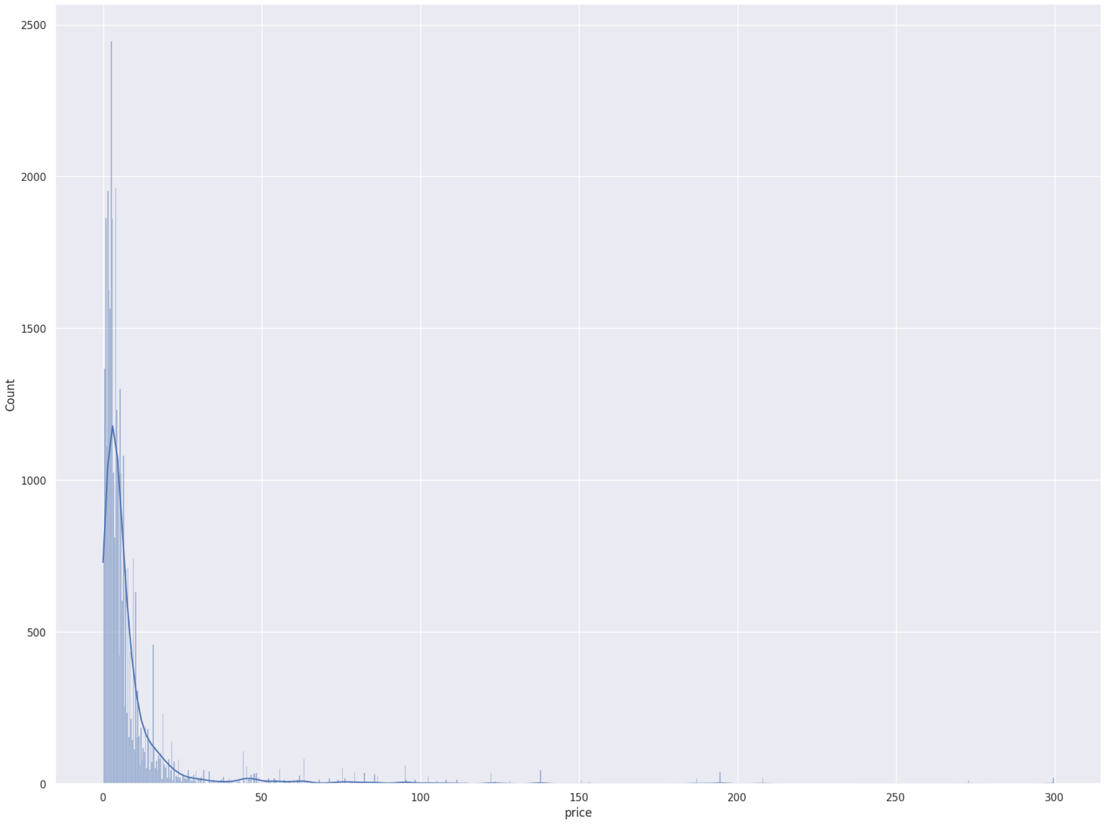
Data

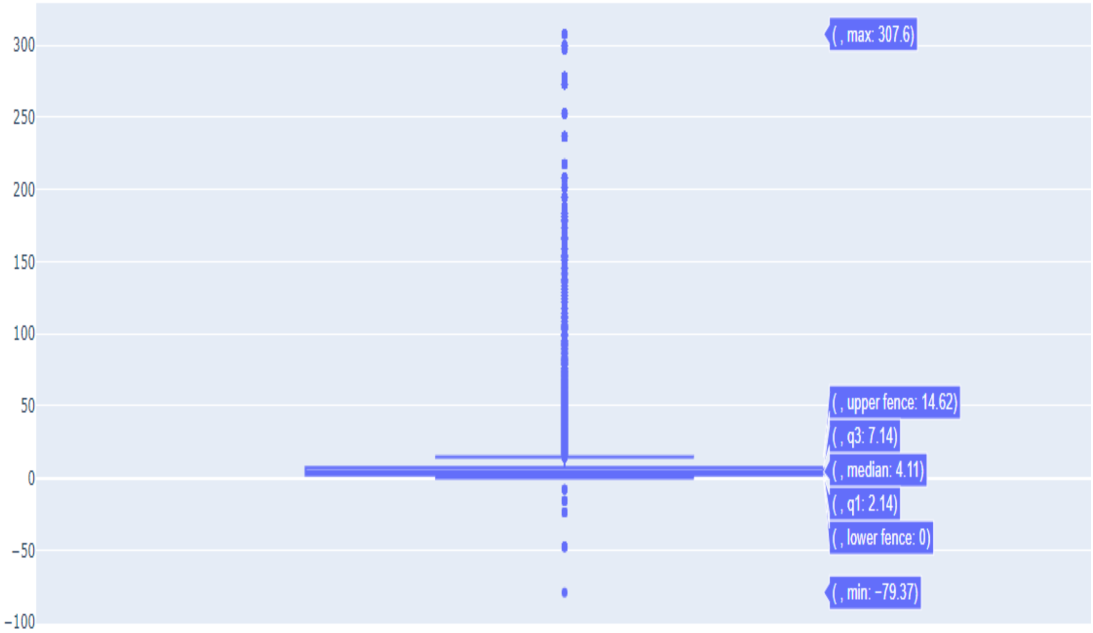
Information

Insights

Decision

Implementation


Count
2500
2000
1500
1000
500
0
0
50
100
150
200
250
300
price


300
250
200
150
100
50
0
−50
−100
(, max: 307.6)
(, upper fence: 14.62)
(, q3: 7.14)
(, median: 4.11)
(, q1: 2.14)
(, lower fence: 0)
(, min: −79.37)